

PREDICTION OF HOUSE PRICING USING MACHINE LEARNING WITH PYTHON

¹Jadhav Vaibhav B., ²Pathak Mandar N., ³Shahare Vivek M.

Computer Engineering,
Loknete Gopinathji Munde Institute of Engineering Education & Research, Nashik

Abstract: This paper offers a top level view about a way to predict house charges using one of a kind regression techniques with the help of python libraries. The proposed approach considered the more subtle aspects used for the calculation of residence fee and offer the greater accurate prediction. It also gives a brief approximately various graphical and numerical techniques with the intention to be required to be expecting the price of a residence. This paper includes what and how the residence pricing model works with the help of machine getting to know and which dataset is used in our proposed version.

Keywords: Machine getting to know, Regression Technique, Classification Technique, Cross validation Technique, K-manner.

INTRODUCTION

House/Home are a basic necessity for someone and their charges vary from region to place primarily based on the facilities available like parking area, locality, and so on. The house pricing is a factor that worries a ton of residents whether or not wealthy or white collar magnificence as you'll be able to by no means choose or gauge the valuing of a residence based on location or offices on hand. Buying of a residence is one of the greatest and good sized choice of a own family as it expends the entirety in their investment finances and every so often covers them beneath loans. It is the hard challenge to predict the accurate values of residence pricing. Our proposed model would make it possible to be expecting the exact expenses of homes.

1.1 Objective

This project is proposed to predict house prices and to get better and accurate results. The stacking algorithm is applied on various regression algorithms to see which algorithm has the most accurate and precise results. This would be of great help to the people because the house pricing is a topic that concerns a lot of citizens whether rich or middle class as one can never judge or estimate the pricing of a house on the basis of locality or facilities available. To accomplish this task, the python programming language is used. Python is a high level programming language for general purpose programming. It enables clear programming on both small and large scales. It is an easily readable language.

1.2 Machine Learning

Machine Learning is a field of Artificial Intelligence which enables PC frameworks to learn and improve in execution with the assistance of information. It is used to study the construction of algorithms that make predictions on data. Machine learning is used to perform a lot of computing tasks. It is also used to make predictions with the use of computers. Machine learning is sometimes also used to devise complex models. The principle point of machine learning is to permit the PCs to learn things naturally without the assistance of people. Machine learning is very useful and is widely used around the whole world. The process of machine learning involves providing data and then training the computers by building machine learning models with the help of various algorithms. Machine learning can be used to make various applications such as face detection application, etc. Machine Learning is a field in software engineering that has changed the way of examining information colossally.

1.3. Python

Python is an elevated level programming language for broadly useful programming. It was created by Guido Van Rossum and released in 1991. It enables clear programming on both small and large scales. Python bolsters various programming standards including object arranged, useful and procedural. Python is an easily readable language. It uses english keywords whereas other programming languages use punctuations. Python utilizes whitespace space as opposed to wavy sections to delimit squares. Python was mainly developed to read codes easily. Python supports various libraries such as Pandas, NumPy, SciPy, Matplotlib etc. It supports various packages such as Xlsx Writer and Xl Rd. Python is an exceptionally helpful language for web improvement and programming advancement. It tends to be utilized to make web applications. It very well may be utilized to peruse and alter documents. It very well may be used to perform complex science. Python has gotten a very well-known language since it can chip away at various stages. Python code can be executed when it is composed. Python is a very significant language since the program is updated without investing additional exertion and energy. Python

bolsters many working frameworks.

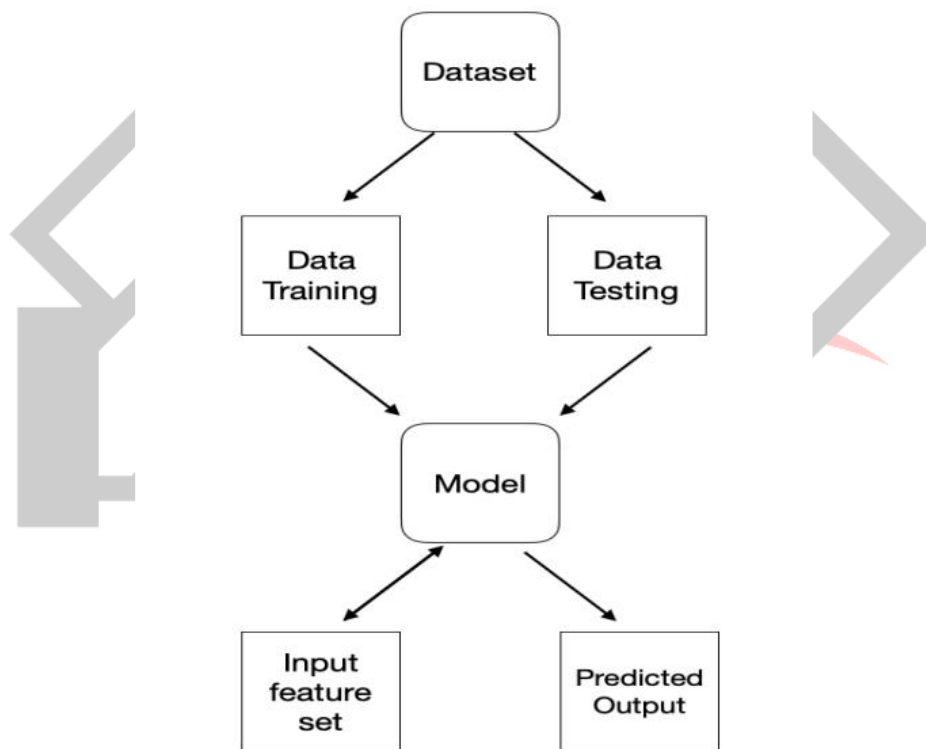
LITERATURE REVIEW

There are a couple of components that impact house costs. In this exploration, part ition these components into three essential get-togethers, there are state of being, thought and territory [2]. States of being are properties constrained by a house that can be seen by human recognizes, including the range of the house, the amount of rooms, the availability of kitchen and parking space, the openness of the yard nursery, the zone of land and structures, and the age of the house [3], while the thought is an idea offered by architects who can pull in potential buyers, for instance, the possibility of a moderate home, strong and green condition, and world class condition. Zone is a critical factor in shaping the expense of a hous e. This is in light of the fact that the zone chooses the normal land cost [4]. Besides, the territory furthermore chooses the basic passage to open workplaces, for instance, schools, grounds, crisis facilit ies and prosperity centers, similarly as family preoccupation workplaces, for instance, strip malls, culinary visits, or much offer awesome landscape [5].

METHODOLOGY

3.1. Data Collection

Data collection is the process of gathering information on variables in a systematic manner. This helps in finding answers too many questions, hypothesis and evaluate outcomes. Data collection is the way toward social event and estimating data on focused factors in a built up framework, which at that point empowers one to address pertinent inquiries and assess results. Information assortment is a part of research in all fields of study including physical and sociologies, humanities and business. While strategies differ by discipline, the accentuation on guaranteeing precise and legitimate assortment continues as before. It has been attempted for various datasets on Kaggle, which would suite our project objective. After looking at a lot of datasets, this dataset is found. It is a house pricing dataset in the city of Ames. This dataset is a very popular machine learning dataset with less scope of errors and variations.



3.2 Data Visualization

Data Visualization is the pictorial or graphical representation of information..It enables to grasp difficult concepts or identify new patterns. Data Visualizat ionis seen by numerous orders as a cutting edge likeness visual correspondence. It includes the creation and investigation of the visual portrayal of information. To impart data plainly and effectively, informat ion representation utilizes measurable illustrations, plots, data designs and different apparatuses.. Effective visualization assists customerswith separating and reason about data and verificat ion. It makes complex data progressively accessible, reasonable and usable. Customers may have expl icit logical endeavors, for instance, making assessments or getting causality, likewise, the structure standard of the reasonable (i.e., indicat ing examinations or demonstrating causality) follows the undertaking. Data Visualization is both a craftsmanship and a science. It is viewed as a piece of particular estimat ions by a couple, yet what's more as a grounded theory improvement device by others. Extended proportions of data made by Web activity and an expanding number of sensors in the earth are suggested as "enormous data" or Web of things. Dealing with, analyzing and passing on this data present good and orderly challenges for data portrayal. The field of data science and experts called data scientists help address this test.

3.3. Data Pre-Processing

It is the process of transforming data before feeding it into the algorithm. It is utilized to change over crude information into a clean dataset. It is an information mining strategy that includes moving crude information into a justifiable organization. The result of data preprocessing is the last dataset utilized for preparing and testing reason. Data preprocessing is an information mining procedure which is utilized to change the crude information in a helpful and productive format. In any Machine Learning procedure, Data Preprocessing is that Encoded, to carry it to such an express, that now the machine can without much of a stretch parse it. Pre-dealing with insinuates the progressions applied to our data before dealing with it to the estimation. Data Preprocessing is a system that is used to change over the rough data into an ideal enlightening assortment. In a manner of speaking, at whatever point the data is amassed from different sources it is assembled in rough setup which isn't feasible for the examination. Genuine information for the most part contains clamors, missing qualities, and perhaps in an unusable organization which can't be legitimately utilized for Machine Learning models. Data preprocessing is required errands for cleaning the information and making it appropriate for an Machine Learning model which likewise expands the precision and proficiency of a Machine Learning model.

Data Cleaning

Data cleaning is the process of detecting and removing errors to increase the value of data. Data cleaning is carried out with the help of data wrangling tools. It is the way toward identifying and amending off base records from a record set, table or database. It finds the deficient information and replaces the messy information. The information is changed to ensure it is exact and right. Information cleaning is the way toward distinguishing and revising mistaken records from a record set, table or database. It is the way toward recognizing inadequate information and afterward supplanting the messy information. The information is changed to ensure that it is exact and right. It is utilized to make a dataset predictable. The principle objective of data cleaning is to distinguish and expel blunders to build the estimation of information in dynamic. The primary center ought to be on distinguishing the right qualities and discover interfaces between different information ancient rarities, for example, patterns and records.

Cross Validation:

Cross validation is a strategy wherein our model is trained using the subset of the dataset and a short time later survey using the basic subset of the dataset. In validation, training is performed on 50% of the dataset and the rest 50% is used for testing purpose. The significant downside of approval strategy is that when it has been prepared for half of the dataset, it might be conceivable that the staying half may contain some helpful data which might be forgotten about at the hour of preparing the model.

K-Fold Cross Validation:

This technique splits the dataset into n number of subsets.

At that point, it has been attempted for preparing on the entirety of the subsets however leave one (k-1) subset for the assessment of the prepared model. This strategy emphasizes k times with an alternate subset turned around for the preparation reason each time.

4. RESULT

To achieve the results, various data mining techniques are utilized in python language. Various factors which affects the house pricing are considered and further worked upon them. Machine learning has been considered to complete out the desired task. Firstly, data collection is performed. Then data cleaning is performed to remove all the errors from the data and make it clean. Then data pre-processing is done. Then with the help of data visualization, different plots are created, which intends to depict the distribution of data in different forms. Towards the end, the business costs of the houses were determined with exactness and accuracy. This could be achieved because a simple stacking algorithm is used to improve the accuracies of the various regression algorithms that are applied on our house pricing dataset so that they would provide better results. Apart from using the regression algorithms, some classification algorithms such as SVM algorithm, decision tree algorithm, Random Forest classifier etc. are taken into consideration and applied on our house pricing dataset. Manner which would help the people to buy houses at a reasonable cost that falls within their budgets

CONCLUSION

The income rate for the houses are calculated the usage of exceptional algorithms. The income fees have been calculated with higher accuracy and precision. This might be of high-quality assist for the humans. To acquire those consequences, diverse statistics mining strategies are applied in python language. The various factors which affect the house pricing need to be considered and paintings upon them. Machine learning has assisted to complete out task. Firstly, the records series is executed. Then data cleaning is completed to get rid of all the mistakes from the facts and make it smooth. Then the information pre-processing is executed. Then with assist of records visualization, one-of-a-kind plots are created. This has depicted the distribution of data in distinctive paperwork. Further, the practise and checking out of the version are executed. It has been located that a number of the classification algorithms had been implemented on our dataset whilst a few had been not. So, the ones algorithms which were not being carried out on our house pricing dataset are dropped and attempted to enhance the accuracy and precision of those algorithms which had been being implemented on our residence pricing dataset. To enhance the accuracy of our type algorithms, a separate stacking algorithm is proposed. It is extraordinarily essential to enhance the accuracy and precision of the algorithms in order to obtain higher consequences. If the outcomes are not accurate then they could be of no assist to the human beings in predicting the income costs of homes. It also made use of facts visualization to gain higher accuracy and results. The income price is calculated for the houses using special algorithms. The income expenses had been calculated with better accuracy and precision. This could

be of first-rate assist for the people.

REFERENCES

- [1] Jain, N., Kalra, P., & Mehrotra, D. (2019). Analysis of Factors Affecting Infant Mortality Rate Using Decision Tree in R Language. In *Soft Computing: Theories and Applications* (pp. 639-646). Springer, Singapore.
- [2] R. A. Rahadi, S. K. Wiryono, D. P. Koesrindartotoor, and I.B. Syamwil, —Factors influencing the price of housing in Indonesia, *Int. J. Hous. Mark. Anal.*, vol. 8, no. 2, pp. 169–188, 2015
- [3] V. Limsombunchai, —House price prediction: Hedonic price model vs. artificial neural network, *Am. J. ...*, 2004
- [4] Kadir, T., & Gleeson, F. (2018). Lung cancer prediction using machine learning and advanced imaging techniques. *Translational Lung Cancer Research*, 7(3), 304-312.
- [5] Liu, J., Ye, Y., Shen, C., Wang, Y., & Erdélyi, R. (2018). A New Tool for CME Arrival Time Prediction using Machine Learning Algorithms: CATPUMA. *The Astrophysical Journal*, 855(2), 109.
- [6] Velankar, S., Valecha, S., & Maji, S. (2018, February). Bitcoin price prediction using machine learning. In *Advanced Communication Technology (ICACT), 2018 20th International Conference on* (pp. 144-147). IEEE.
- [7] Malhotra, R., & Sharma, A. (2018). Analyzing Machine Learning Techniques for Fault Prediction Using Web Applications. *Journal of Information Processing Systems*, 14(3).
- [8] Choo, M. S., Uhm, S., Kim, J. K., Han, J. H., Kim, D. H., Kim, J., & Lee, S. H. (2018). A Prediction Model Using Machine Learning Algorithm for Assessing Stone-Free Status after Single Session Shock Wave Lithotripsy to Treat Ureteral Stones. *The Journal of urology*.
- [9] Nilashi, M., Ibrahim, O., Ahmadi, H., Shahmoradi, L., & Farahmand, M. (2018). A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques. *Biocybernetics and Biomedical Engineering*, 38(1), 1-15.
- [10] Fan, C., Cui, Z., & Zhong, X. (2018, February). House Prices Prediction with Machine Learning Algorithms. In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing* (pp. 6-10). ACM.